Original article

Journal of High School Science

# Beyond normality: comparative tail-risk analysis of S&P 500 returns

Chen C

Submitted: April 1, 2026, Revised: version 1, May 20, 2026, version 2, May 26, 2026, version 3, May 27, 2026
Accepted: May 31, 2026

**Abstract**
Financial models normally assume that investment returns follow a normal distribution because of the latter's mathematical simplicity and analytical convenience. However, real market returns often exhibit non-normal characteristics such as heavy tails and skewness, which can lead to underestimation of financial risk. In this study, 31 years of SPY daily return data were analyzed to evaluate the distributional behavior and tail risk of S&P 500 returns. Descriptive statistics, distribution fitting, Quantile-Quantile (Q-Q) plots, and formal normality tests were used to examine deviations of the SPY returns from normality. Additionally, a heavy-tailed t distribution, a two-component Gaussian Mixture Model (GMM-2), and Extreme Value Theory (EVT) were evaluated for tail-risk modeling. Model performance was further assessed using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Value at Risk (VaR), Expected Shortfall (ES), exceedance backtesting, and rolling-window calibration analyses across different market environments. The results show that, compared to the normal distribution, SPY returns exhibit negative skewness, pronounced central peaks, and heavy tails. The normal model performs reasonably well in moderate-tail regions and during stable market conditions, but increasingly underestimates downside risk at deeper confidence levels and during stressed market periods. Among the models, the t-distribution provides the strongest overall balance between model simplicity and tail-risk representation, while EVT performs best for rare and extreme loss events. Rolling-window and annual analyses further show that model performance varies across different market regimes and volatility conditions. Although heavy-tailed models improve tail-risk calibration relative to the normal distribution, rolling-window analyses show persistent exceedance underestimation across all models, suggesting that time-varying volatility and regime-dependent behavior play a major role in financial tail-risk dynamics. Overall, the findings highlight the limitations of normal-distribution assumptions in financial risk modeling and demonstrate that tail-risk behavior varies substantially across different market regimes, volatility conditions, and tail depths. No single model consistently provides the best performance under all conditions, underscoring the importance of appropriate model selection in financial risk estimation.

**Keywords**
Tail risk, Extreme Value Theory, Value at Risk, Expected Shortfall, Heavy tails, Financial markets, Risk forecasting, Backtesting, Volatility modeling, Market regimes

Charles Chen, James E. Taylor High School, 20700 Kingsland Blvd, Katy, TX 77450, USA. chfchen08@gmail.com


https://doi.org/10.64336/001c.162826

## 1. Introduction

More than half of American households hold investments in stocks, mutual funds, or other financial securities (1). Over the long term, returns from these assets generally exceed those from traditional bank savings (2). To evaluate investment performance and manage financial risk, investors and institutions commonly use statistical and mathematical models to analyze return behavior. Many classical financial models, including the Capital Asset Pricing Model (CAPM) and Value at Risk (VaR) model, are based on the assumption that returns or log-returns follow a normal distribution (3,4). The normal distribution is mathematically convenient because of its symmetric bell shape and relatively simple analytical properties.

However, real financial markets are influenced by economic events, news, investor behavior, and market uncertainty, causing return distributions to deviate from normal assumptions. Fama found that stock returns often exhibit heavy tails, meaning extreme market movements occur more frequently than predicted by a normal distribution (5). Cont further showed that these non-normal characteristics can cause financial models to underestimate market risk, particularly during periods of market stress (6). Previous studies have also shown that although return distributions tend to move toward normality over longer time horizons, significant tail-risk behavior can still persist because of major financial crises and volatility clustering (5,6).

Because of these limitations, alternative approaches such as heavy-tailed distributions, Extreme Value Theory (EVT), and rolling-window risk estimation methods have been widely used to improve tail-risk modeling and downside-risk forecasting under changing market conditions (7-9). In particular, EVT has been commonly applied to model rare and extreme financial losses, while rolling-window and backtesting approaches are frequently used to evaluate the stability and predictive performance of risk models over time (8,9).

These observations motivate the need to evaluate whether alternative statistical models can provide improved representation of SPY return behavior and tail risk. In this study, 31 years of SPY daily return data were analyzed using descriptive statistics, distribution fitting, Q-Q plots, formal normality tests, and multiple tail-risk models, including the t-distribution, Gaussian Mixture Model (GMM-2), and Extreme Value Theory (EVT). In addition, Value at Risk (VaR), Expected Shortfall (ES), exceedance backtesting, annual calibration analysis, and rolling-window analysis were used to evaluate practical tail-risk estimation performance under different market conditions.

The study aims to evaluate how different statistical models represent tail risk and to examine the limitations of traditional normal-distribution assumptions in financial risk estimation.

## 2. Materials and Methods

### **2.1** SPY returns

The historical price data for the S&P 500 ETF (SPY) were obtained from Yahoo Finance. The study period was from January 1, 1994, to December 31, 2024. The adjusted close price was used to calculate returns as it provides a more accurate measure of total returns. Log-returns were used in this study because they are additive over time and commonly employed in financial modeling. Since daily returns are generally small, log-returns are numerically very close to simple returns. Throughout this study, the term "returns" refers specifically to log-returns. The equations utilized were, $R_t = \frac{P_t}{P_{t-1}} - 1$ (Eq1), $r_t = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln\left(1 + R_t\right)$ (Eq 2) and $r_T = \ln\left(\frac{P_T}{P_0}\right) = \sum_{t=1}^{T} \ln \frac{P_t}{P_{t-1}} = \sum_{t=1}^{T} r_t$ (Eq 3), where $P_t$ is the price of an investment at time (day) $t$, $R_t$ is the simple return, $r_t$ is the log-return, $T$ is the holding period in days, and $r_T$ is the cumulative log-return over period $T$.

**2.2** Descriptive statistics

Maximum, minimum, mean, standard deviation, skewness, and kurtosis were calculated for SPY returns.

**2.3** Distribution and Q-Q plots

SPY return histograms were plotted on a probability-density scale using 50 bins together with smoothed Kernel Density Estimation (KDE) curves. Quantile-Quantile (Q-Q) plots were used to compare empirical return quantiles with theoretical distribution quantiles.

**2.4** Normality testing

Two statistical tests were used to check whether SPY returns were normally distributed. The Jarque–Bera (JB) test checks whether a dataset's skewness and kurtosis are consistent with those of a normal distribution. The Anderson-Darling (AD) test measures how well a dataset fits a specified theoretical distribution.

**2.5** Gaussian Mixture Model

A two-component Gaussian Mixture Model (GMM-2) was used to represent returns as a weighted combination of two normal distributions, allowing simultaneous modeling of the central high-density region and higher-volatility tail behavior. Although the two GMM components may loosely resemble low- and high-volatility market states, they should be interpreted as statistical mixture components rather than directly observed economic regimes.

**2.6** t-distribution model

The t-distribution was used as a heavy-tailed alternative to the normal distribution for modeling extreme return behavior.

**2.7** Extreme Value Theory

Extreme Value Theory (EVT) was used to analyze rare and extreme downside return events. In this study, the daily returns were sorted from lowest to highest, and the lowest 5% of returns were selected for tail analysis. This threshold was chosen to balance statistical stability with emphasis on the extreme-tail region of the return distribution.

A Weibull distribution was then fitted using linear regression based on the logarithmic relationship between exceedance probability and normalized extreme returns. The EVT approach was included because it specifically emphasizes extreme-tail behavior beyond the range typically captured by standard parametric distributions such as the normal distribution.

**2.8** Model selection using AIC and BIC

To evaluate which theoretical distribution model best fits the return data, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used. Both criteria evaluate the trade-off between goodness-of-fit and the number of model parameters, with lower values indicating a better balance between model accuracy and complexity.

**2.9** Value at Risk and Expected Shortfall
Value at Risk (VaR) and Expected Shortfall (ES) were used to evaluate and compare tail risk across different distribution models. VaR estimates the maximum expected loss over a specified time horizon at a given confidence level, while ES measures the average loss of all returns exceeding the VaR threshold.

**2.10** VaR exceedance backtesting
VaR exceedance backtesting was performed to evaluate how accurately each model predicted the frequency of extreme losses. For a given confidence level, the observed number of returns below the predicted VaR threshold was compared with the theoretically expected number of exceedances. A ratio of observed-to-expected exceedances close to 1 indicates good calibration, while ratios significantly above or below 1 suggest underestimation or overestimation of tail risk, respectively.

**2.11** Rolling-window analysis
A rolling-window analysis was used to evaluate the dynamic predictive performance of the tail-risk models. Using a 500-day rolling window, model parameters and VaR estimates were recalculated sequentially based only on the most recent historical returns, and the resulting predictions were compared against the subsequent realized return. This approach allowed assessment of out-of-sample tail-risk forecasting performance and captured changing market conditions. Additionally, a sensitivity analysis using 250-day and 750-day rolling-window lengths was conducted to evaluate the robustness of the rolling-window backtesting results to the selected window length.

To further examine the time-dependent behavior of rolling-window VaR calibration, annual observed-to-expected exceedance ratios were calculated from the 500-day rolling-window backtesting results. This analysis was used to evaluate how model calibration changed across different market environments and volatility regimes over time.

**2.12** Analysis tools
All calculations and plots were performed in Python using data science libraries, including Pandas and NumPy for data handling, Matplotlib for graphs, SciPy and Statsmodels for distribution models and statistical testing, and Scikit-learn for the Gaussian Mixture Model.

## 3. Results and Discussion

**3.1** Descriptive statistics of returns
To provide an overview of SPY performance, descriptive statistics of returns for three different time horizons – daily, monthly, and annual – were calculated as shown in Table 1.

For daily returns, the dataset contains 7,804 observations ranging from -11.59% to 13.56%, with a near-zero mean return of 0.04% and a standard deviation of 1.19%. The mean returns and standard deviations across daily, monthly, and annual horizons are broadly consistent with the theoretical scaling behavior of log-returns: mean returns increase approximately proportionally with time horizon, while standard deviation increases more slowly, ~ in proportion to the square root of time.

**Table 1.** Descriptive statistics of SPY returns across daily, monthly and annual time horizons

| Parameter | Observations | Maximum | Mean | Minimum | Standard Deviation | Skewness | Excess Kurtosis |
|---|---|---|---|---|---|---|---|
| Daily | 7,804 | 13.56% | 0.04% | -11.59% | 1.19% | -0.29 | 10.84 |
| Monthly | 372 | 11.95% | 0.83% | -18.04% | 4.36% | -0.75 | 1.28 |
| Annual | 31 | 32.23% | 9.97% | -45.88% | 17.95% | -1.34 | 1.87 |

Daily returns exhibit a small negative skewness (-0.29), indicating that large losses occurred slightly more frequently than large gains. Monthly and annual returns exhibit even more negative skewness values, suggesting that longer investment horizons remain influenced by major market downturns despite aggregation effects predicted by the Central Limit Theorem (10).

Daily returns also exhibit very high excess kurtosis (10.84), reflecting a sharp central peak and heavy tails relative to a normal distribution. This indicates that extreme return events occur far more frequently than predicted under normal assumptions. In contrast, monthly and annual returns exhibit substantially smaller excess kurtosis values, consistent with the tendency toward normality predicted by the Central Limit Theorem (5).

Because tail-risk estimation and risk backtesting are primarily associated with higher-frequency return behavior, the remaining analyses in this study focus mainly on daily return data.

**3.2** Distributional characteristics and model fitting

Multiple theoretical distribution models were employed to fit the historical daily return data, particularly to the tails. While the normal distribution is the traditional benchmark in financial modeling, it often underestimates tail risk. Therefore, the heavy-tailed t-distribution and a two-component Gaussian Mixture Model (GMM-2) were also evaluated to examine whether they provided improved representation of heavy-tailed behavior. Figure 1 compares the distribution of SPY daily returns with the theoretical models, and Figure 2 presents a zoomed-in comparison of the left tail. Figure 3 shows the Q-Q plots for all three selected theoretical models, with the 45° diagonal reference line ($y = x$) representing a perfect agreement between empirical and theoretical distributions, and larger deviations from the line indicating poorer distributional fit.

For the t-distribution, the estimated parameters are were, Location $\mu = 0.08\%$, Scale $\sigma = 0.71\%$, Degrees of Freedom $\nu = 2.79$ (heavy tails)

For the two-component GMM, the fitted parameters were,
Component 1 - dominant low-volatility component: Weight $w = 77\%$, Mean $\mu = +0.11\%$, Standard Deviation $\sigma = 0.71\%$
Component 2 - small high-volatility component: Weight $w = 23\%$, Mean $\mu = -0.21\%$, Standard Deviation $\sigma = 2.10\%$

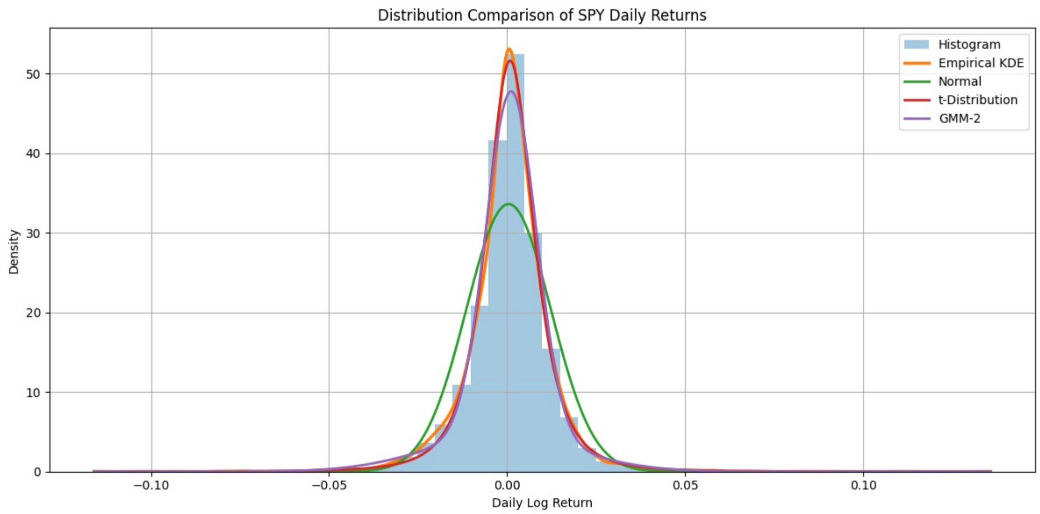


**Figure 1.** Comparison of the empirical SPY daily return distribution with theoretical models

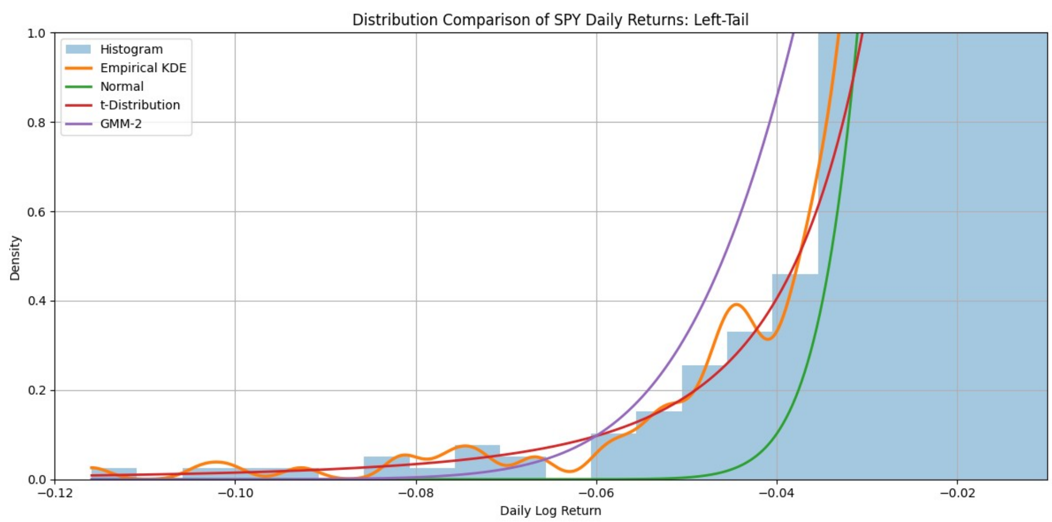


**Figure 2.** Zoomed-in comparison of the left-tail region of SPY daily return distribution with theoretical models

The figures show that:

[1] The empirical SPY daily return distribution is bell-shaped. However, it exhibits a noticeably sharp central peak and heavy tails.

[2] In the main-body region, the normal distribution shows a lower central peak and a wider body than the empirical distribution, indicating that the normal assumption allocates excessive probability to the moderate return region. The GMM-2 model provides a noticeable improvement, although its peak remains slightly below the empirical curve. Among the three models, the t-distribution provides the closest fit, with a central peak nearly identical to the empirical data.

[3] For the left-tail region, the normal distribution increasingly underestimates the probability of large losses, particularly beyond

approximately -4%. The GMM-2 model improves tail fitting, particularly in the intermediate left-tail region, but still underestimates the most extreme losses below approximately -7%. The t-distribution provides the closest fit to the empirical left tail.

[4] The Q-Q plots further support these findings. The normal distribution Q-Q plot exhibits an S-shaped deviation, indicating underestimation of extreme returns at both sides. The GMM-2 model reduces the magnitude of these deviations but retains a similar overall pattern. In contrast, the t-distribution displays reversed tail deviations, suggesting tails slightly heavier than the observed empirical extremes.

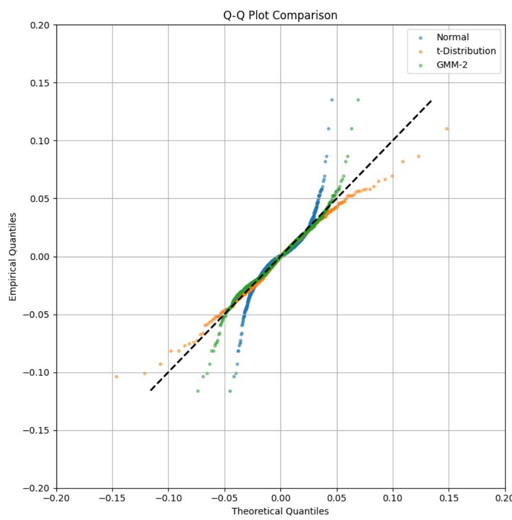


**Figure 3.** Q-Q plots of SPY daily returns relative to theoretical models

Overall, the results show that SPY returns exhibit tail behavior that is intermediate between the thin-tailed normal distribution and the strongly heavy-tailed t-distribution, with the t-distribution providing the closest overall representation among the tested models. Consistent with the earlier descriptive statistics, the findings further confirm that extreme losses occur more frequently than predicted under normal assumptions, while the GMM-2 model only partially captures the heavy-tailed structure of returns. These results highlight the limitations of normal assumptions in financial modeling and motivate the use of heavy-tailed or mixture-based approaches for downside-risk estimation.

**3.3** Formal normality tests

To formally evaluate whether SPY returns followed a normal distribution, two commonly used statistical tests were employed: the Jarque-Bera (JB) test and the Anderson-Darling (AD) test. Both tests strongly rejected the null hypothesis of normality at the 5% significance level. For the JB test, the JB value of 38,293 > chi-square critical value of 6, thereby rejecting normality. For the AD Test, the $A^2$ value of 135 > critical value of 0.8, thereby rejecting normality.

**3.4** Model selection (AIC/BIC)

To compare model performance while accounting for differences in model complexity, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used. Both criteria evaluate the trade-off between goodness-of-fit and the number of model parameters. The relevant equations are, $AIC = 2k - 2\ln(L)$ (Eq 4) and $BIC = k\ln(n) - 2\ln(L)$ (Eq 5), where $k$ is the number of model parameters, $L$ is the maximum likelihood of the fitted model, and $n$ is the sample size.

Lower AIC and BIC values indicate a more favorable balance between statistical fit and model simplicity. Unlike pure goodness-of-fit measures, AIC and BIC penalize excessive model complexity, thereby discouraging overfitting. Model performance comparisons are presented in Table 2.

The results show that the t-distribution has the lowest AIC and BIC values among the tested models, indicating the best overall balance between model fit and complexity. Although the GMM-2 model substantially improves distribution fit relative to the single normal model, its additional complexity did not provide sufficient improvement relative to the t-distribution. These findings suggest that heavy-tailed behavior plays a more dominant role in SPY return non-normality than mixture-distribution structure alone.

**Table 2.** Comparison of AIC and BIC values

| Model | Parameters | Log-Likelihood | AIC | BIC |
|---|---|---|---|---|
| Normal | 2 | 23,533 | -47,062 | -47,048 |
| GMM-2 | 5 | 24,493 | -48,977 | -48,942 |
| t-Distribution | 3 | 24,579 | -49,152 | -49,131 |

**3.5** Comparison of downside risk: VaR and ES

To illustrate the practical implications of non-normality, Value at Risk (VaR) and Expected Shortfall (ES) were compared across multiple confidence levels (95%, 98%, and 99%). In addition to the aforementioned distribution models, Extreme Value Theory (EVT) was also included because it is specifically designed to model rare and extreme tail events, and the lowest 5% returns were used to estimate the Weibull distribution parameters in EVT estimates. Table 3 and Figure 4 compare the calculated VaR and ES values.

**Table 3.** Tail-Risk (VaR and ES) comparison among different distribution models

| Tail Risk | VaR Values | | | ES Values | | |
|---|---|---|---|---|---|---|
| Confidence Level | 95% | 98% | 99% | 95% | 98% | 99% |
| Empirical | -1.86% | -2.64% | -3.29% | -2.86% | -3.88% | -4.82% |
| Normal | -1.91% | -2.40% | -2.72% | -2.41% | -2.83% | -3.12% |
| GMM-2 | -1.90% | -3.06% | -3.80% | -3.06% | -4.03% | -4.66% |
| t-Distribution | -1.64% | -2.50% | -3.33% | -2.84% | -4.12% | -5.38% |
| EVT | -1.40% | -2.90% | -3.82% | -2.87% | -4.08% | -4.84% |

The tail-risk comparison shows that:

[1] For the normal distribution, the 95% VaR estimate is close to the empirical VaR, but it increasingly underestimates downside risk at higher confidence levels, indicating that the normal distribution captures moderate downside risk reasonably well but fails to capture extreme losses. A similar pattern is observed for ES, where the normal model increasingly underestimates ES as the confidence level increases.

[2] The GMM-2 model shows more conservatism at deeper confidence levels. While its 95% VaR estimate is close to the empirical result, the model produces increasingly conservative VaR estimates at higher confidence levels. The ES estimates from GMM-2 are generally close to the empirical values, indicating improved representation of tail risk relative to the normal distribution.

[3] The t-distribution provides significantly improved extreme-tail estimation relative to the normal model. Although the 95% VaR slightly underestimates the empirical value, the 98% and 99% VaR values closely match the empirical estimate. Similarly, the 95% ES value is nearly identical to the empirical ES. However, at the 99% confidence level, the t-distribution produces more extreme ES estimates than the empirical result, indicating that the fitted t-distribution may overestimate the severity of the most extreme tail events. This is consistent with the Q-Q plot results, where the t-distribution exhibited tails slightly heavier than the empirical data.

[4] The EVT model exhibits the strongest emphasis on extreme-tail behavior. At the 95% confidence level, EVT produces the smallest-magnitude VaR estimate, reflecting the fact that EVT is not intended for moderate-tail modeling. However, at the 98% and 99% levels, EVT produces some of the most conservative VaR estimates closely matching or exceeding the empirical tail risk. The EVT ES estimates align closely with the empirical ES values. These results suggest that EVT is highly effective for modeling rare and extreme losses but less appropriate for moderate-risk regions.

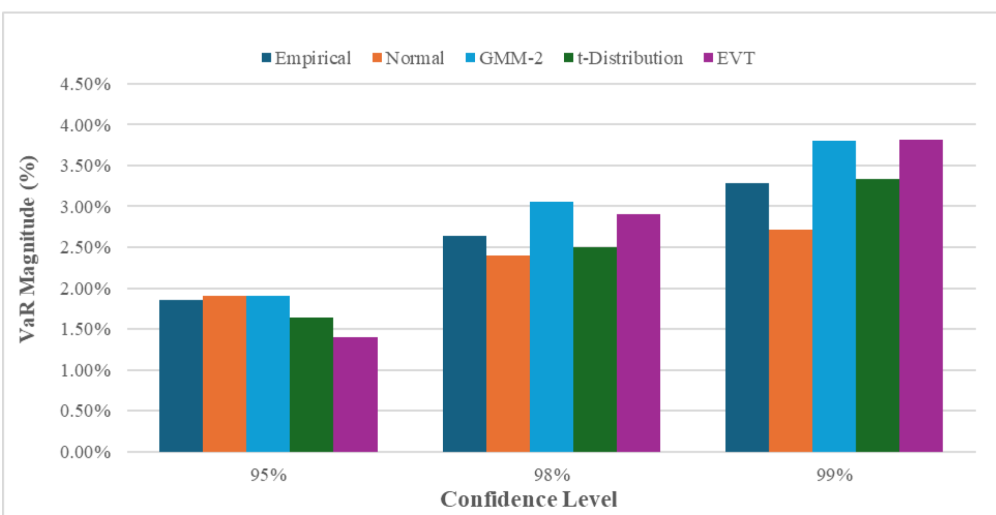


**Figure 4.** Comparison of VaR estimates across distribution models and confidence levels

Overall, model differences become increasingly pronounced at deeper confidence levels. While the normal distribution remains reasonably close to empirical risk estimates at moderate confidence levels, the heavy-tailed models provide improved representation of rare market losses, although some models become increasingly conservative in the deepest tails.

**3.6** VaR exceedance backtesting

While VaR and ES comparisons evaluate the magnitude of predicted tail risk, they do not directly measure how accurately the models predict the observed frequency of extreme losses. Therefore, VaR exceedance backtesting was conducted to evaluate the practical predictive performance of each model. The expected and observed exceedances are defined as, $N_{expected} = (1 - c) \times n$ (Eq 6) where $N_{observed}$ = Observed number of returns below predicted VaR threshold (Eq 7), $c$ is the confidence level, $n$ is the total number of observations. The exceedance ratio is calculated as, $R = \frac{N_{observed}}{N_{expected}}$ (Eq 8). A ratio close to 1 indicates good agreement between expected and observed exceedances. A ratio greater than 1 indicates underestimation of downside risk by the model, while a ratio less than 1 indicates overestimation.

**Table 4.** VaR exceedance backtesting Results. Comparison of observed-to-expected VaR exceedance ratios across different confidence levels and distribution models

| $N_{observed}/N_{expected}$ | Confidence Level | | |
|---|---|---|---|
| Model | 95% | 98% | 99% |
| Normal | 0.95 | 1.33 | 1.83 |
| GMM-2 | 0.96 | 0.62 | 0.67 |
| t-Distribution | 1.27 | 1.19 | 0.96 |
| EVT | 1.65 | 0.79 | 0.65 |

The VaR exceedance backtesting results indicate that

[1] The normal distribution performs reasonably well at the 95% confidence level but calibration deteriorates substantially at deeper confidence levels.

[2] The GMM-2 model produces exceedance ratios below 1 at all confidence levels, indicating systematic overestimation of tail risk.

[3] The t-distribution produces exceedance ratios relatively close to 1 across all confidence levels, particularly at deeper confidence levels, indicating generally strong tail-risk calibration.

[4] The EVT model shows the greatest overestimation at the deepest confidence levels, reflecting its strong emphasis on extreme-tail behavior.

Overall, in alignment with the earlier distribution-fitting and tail-risk analyses, model differences become increasingly pronounced at deeper confidence levels. The normal distribution shows progressively weaker calibration for extreme-loss frequency, while the heavy-tailed models provide improved exceedance behavior. Among the tested parametric models, the t-distribution produces the most balanced overall exceedance calibration across all confidence levels.

**3.7** *Annual VaR analysis*
Previous analyses evaluated overall tail-risk behavior across the full 31-year dataset; however, market conditions and return distributions can vary substantially over time. Therefore, an annual VaR analysis was conducted to examine how tail-risk estimation varies across different market regimes.

For each year, VaR estimates from the empirical model, normal distribution, and t-distribution were calculated at the 95% and 98% confidence levels. Model calibration was evaluated using the ratio between model-estimated VaR and empirical VaR as, $\text{VaR Ratio} = \frac{|\ Model\ \ VaR\ |}{|\ Empirical\ \ VaR\ |}$ (Eq 9). A ratio below 0.9 was classified as risk underestimation, a ratio between 0.9 and 1.1 was considered acceptable, and a ratio above 1.1 was classified as risk overestimation.

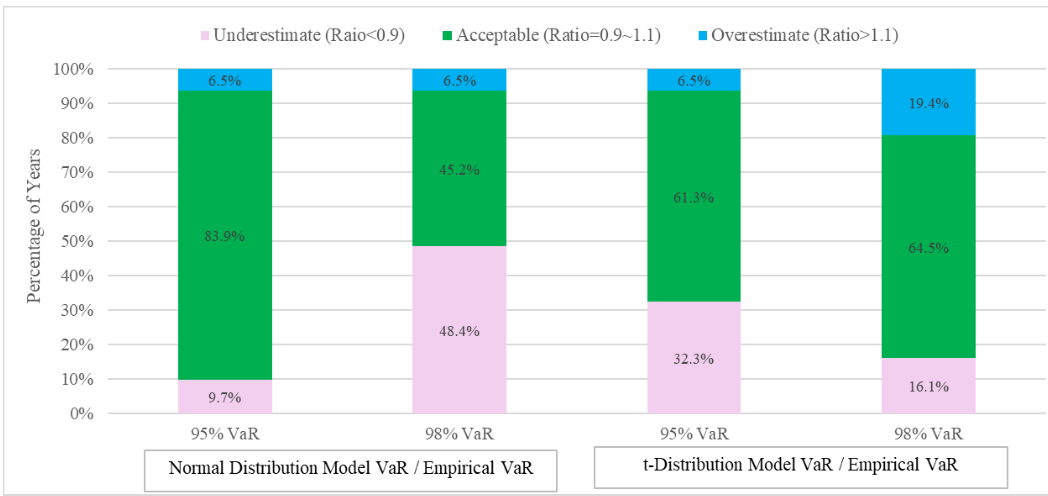


**Figure 5.** Annual VaR calibration across confidence levels and distributional Models. Comparison of percentage of years in which the model-to-empirical VaR ratio falls into three calibration ranges for the normal and t-distribution models

Figure 5 shows the annual VaR calibration across confidence levels and distributional models. At the 95% confidence level, the normal distribution performs reasonably well, with 83.9% of years falling within the acceptable range and only 9.7% of years showing underestimation of tail risk. However, at the 98% confidence level, the normal model deteriorates substantially, indicating that the normal distribution increasingly fails to capture deeper tail behavior during extreme market conditions. The t-distribution shows a different pattern. At the 95% confidence level, the model shows significantly more frequent underestimation than the normal distribution. However, unlike the normal distribution, the t-distribution improves at the deeper 98% confidence level, where underestimation decreases significantly. This suggests that the heavy-tailed t-distribution becomes increasingly advantageous when modeling more extreme downside risk.

Overall, the annual analysis further demonstrates that the normal distribution may provide acceptable approximation for moderate-tail risk but becomes progressively less reliable for deeper tail estimation. In contrast, the t-distribution provides more stable and accurate calibration for extreme-tail risk across different market environments.

Additional analyses were performed for 1995 and 2008, representing relatively stable low-volatility and stressed high-volatility market environments, respectively.

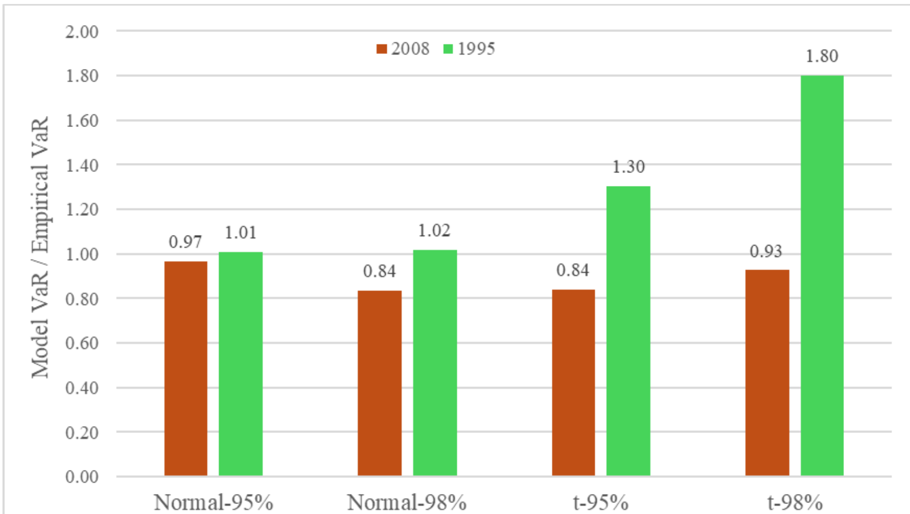


**Figure 6.** Annual VaR Calibration for 1995 and 2008. Comparison of model-to-empirical VaR ratios for the normal and t-distribution models in 1995 and 2008 at the 95% and 98% confidence levels

Results from the stressed year of 2008 show that both the normal and t-distribution models underestimate tail risk for both confidence levels. For the normal distribution, the model-to-empirical VaR ratio decreased from 0.97 at the 95% level to 0.84 at the 98% level, indicating worsening underestimation as tail depth increased. However, the t-distribution improved from 0.84 at 95% to 0.93 at 98%, suggesting better calibration.

In contrast, during the relatively stable year of 1995, the normal distribution remains close to empirical observations at both confidence levels. However, the t-distribution becomes increasingly conservative, particularly in the deeper tails.

These results suggest that model performance depends strongly on market conditions, with the normal distribution performing reasonably well during stable low-volatility periods while heavy-tailed models become more advantageous during stressed market environments.

### **3.8** Rolling-window backtesting analysis

A rolling-window analysis was conducted to evaluate the time-varying predictive performance of different VaR models under changing market conditions. Unlike static full-sample or annual analyses, the rolling-window method used only information available prior to each forecast date, thereby more closely reflecting real-world risk management practice.

A 500-trading-day rolling window (~1.4 trading years) was selected to balance statistical stability and responsiveness to evolving market conditions. For each trading day *t*, the previous 500 daily returns (*t-500* to *t*-1) were used to estimate VaR using historical simulation, normal distribution, and t-distribution methods. The resulting VaR estimate was then compared against the realized return observed on day *t*. If the realized return fell below the predicted VaR threshold, a VaR exceedance (or violation) was recorded.

Over the full sample, the observed number of exceedances was compared with the theoretically expected exceedance frequency implied by the confidence level. For a properly calibrated 99% VaR model, approximately 1% of observations are expected to exceed the VaR threshold. The observed-to-expected exceedance ratio was used to evaluate dynamic tail-risk calibration, with values greater than 1 indicating underestimation of downside risk.

The rolling-window method allows the analysis to capture changing market conditions and tail-risk behavior over time, thereby providing a more realistic assessment of practical risk-model performance. Table 5 summarizes the rolling-window backtesting results, and Figure 7 presents the corresponding time-series comparison.

**Table 5.** Rolling-window VaR backtesting results. Comparison of expected and observed exceedances for the rolling 99% VaR analysis using a 500-day rolling window under the historical, normal, and t-distribution models

| Model | Expected Exceedances | Observed Exceedances | Ratio |
|---|---|---|---|
| Historical | 73 | 116 | 1.59 |
| Normal | 73 | 184 | 2.52 |
| t-Distribution | 73 | 124 | 1.70 |

The rolling-window backtesting results show that all models produced more VaR exceedances than expected, indicating that realized downside risk occurred more frequently than predicted. The normal model showed the largest underestimation, with observed exceedances reaching 2.52 times the expected level. The historical simulation performed best, with an exceedance ratio of 1.59, while the t-distribution reduced underestimation significantly relative to the normal model but remained slightly less accurate than the rolling historical-simulation approach. These results suggest that heavy-tailed modeling improves upon normal assumptions, but recent empirical tail behavior remains particularly important for dynamic risk forecasting.

Although the historical simulation and t-distribution models substantially improved rolling-window calibration relative to the normal distribution, all models continued to underestimate tail risk, yielding exceedance ratios significantly above the expected theoretical value. These results suggest that static distributional assumptions alone are not sufficient to fully capture the dynamic and clustered nature of financial tail risk. In particular, volatility clustering, regime transitions, and rapidly changing market conditions may contribute more strongly to tail-risk forecasting error than the choice of unconditional return distribution itself. More adaptive approaches, such as conditional volatility or regime-switching models, may therefore be necessary for improved dynamic tail-risk forecasting.

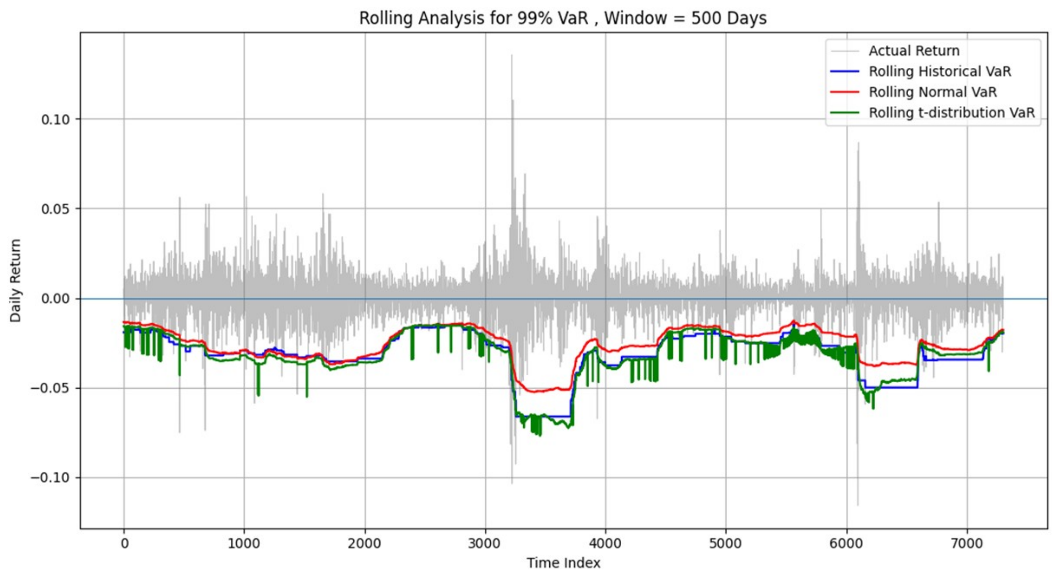


**Figure 7.** Rolling 99% VaR time-series comparison. Time-series comparison of realized SPY daily returns with rolling 99% VaR estimates generated using the historical simulation, normal distribution, and t-distribution models based on a 500-trading-day rolling window

The rolling time-series comparison further supports these results. The rolling t-distribution VaR curve tracks the historical simulation much more closely than the normal-distribution model, particularly during stressed market periods. In contrast, the normal VaR curve remains less conservative, contributing to its larger number of exceedances. The t-distribution also exhibits sharper short-term adjustments, reflecting its greater sensitivity to extreme-return behavior.

To examine the sensitivity of the rolling-window backtesting results to the window length, additional analyses using 250-day and 750-day rolling windows were conducted for 99% VaR, and the comparison is shown in Figure 8.

The sensitivity analysis results show that the observed-to-expected exceedance ratios remain relatively stable across different rolling-window lengths. Both the 250-day and 750-day windows produce slightly lower ratios than the 500-day window, indicating that the relationship between window length and exceedance calibration is not strictly monotonic and that the rolling-window backtesting results are not strongly sensitive to the selected window length. The results also confirm that the normal distribution consistently produces the highest exceedance ratios, while the historical simulation and t-distribution models provide substantially improved calibration across all rolling windows.

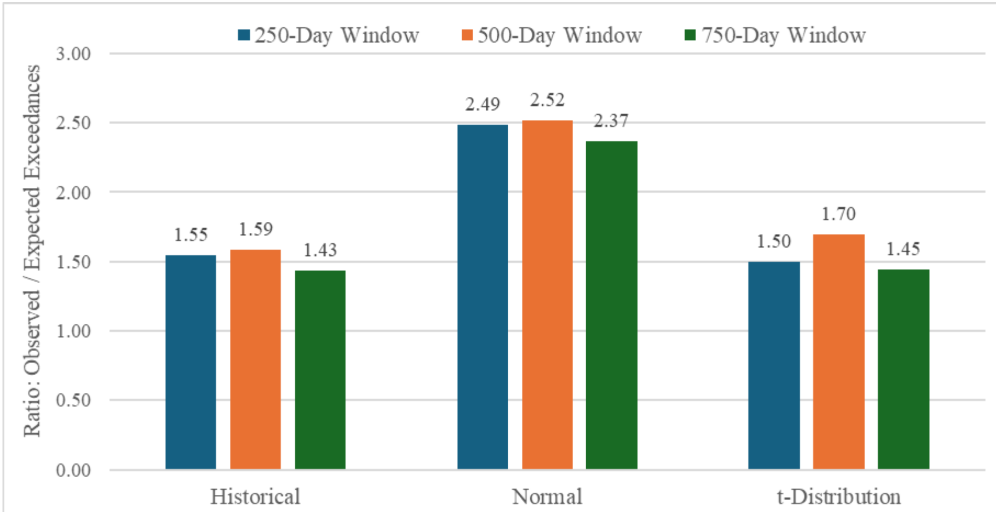


**Figure 8.** Rolling-window sensitivity analysis of 99% VaR exceedance ratios across different window lengths. Comparison of observed-to-expected exceedance ratios. Ratios greater than 1 indicate underestimation of downside risk

**3.9** Annual rolling-window exceedance analysis

To further investigate the regime-dependent behavior of rolling-window VaR calibration, the annual observed-to-expected exceedance ratios from the 500-day rolling-window analysis were examined across different market conditions, and the results are shown in Figure 9.

Based on the annual rolling-window exceedance calculation methodology, the backtesting results depend not only on current-year market performance, but also on the preceding market conditions within the rolling window. For the 500-day rolling-window VaR estimates, return data from the previous two trading years were used. The results show that when the current-year return is substantially lower than either of the preceding two years, exceedance ratios tend to increase, indicating greater underestimation of downside risk. This pattern is observed during stressed periods such as the 2008 financial crisis and the 2020 pandemic-related volatility period. In contrast, when the current-year return is substantially higher than either of the preceding two years, exceedance ratios tend to remain low or near zero, reflecting relatively conservative VaR estimates. This behavior is observed during recovery periods such as 2009–2010 and 2021.

Additionally, the results show that the three models generally follow similar time-dependent patterns, with the primary difference being the magnitude of the exceedance ratios. In general, the normal distribution produces the largest exceedance ratios, indicating more severe underestimation of downside risk. These findings are broadly consistent with the whole-sample annual VaR analysis.

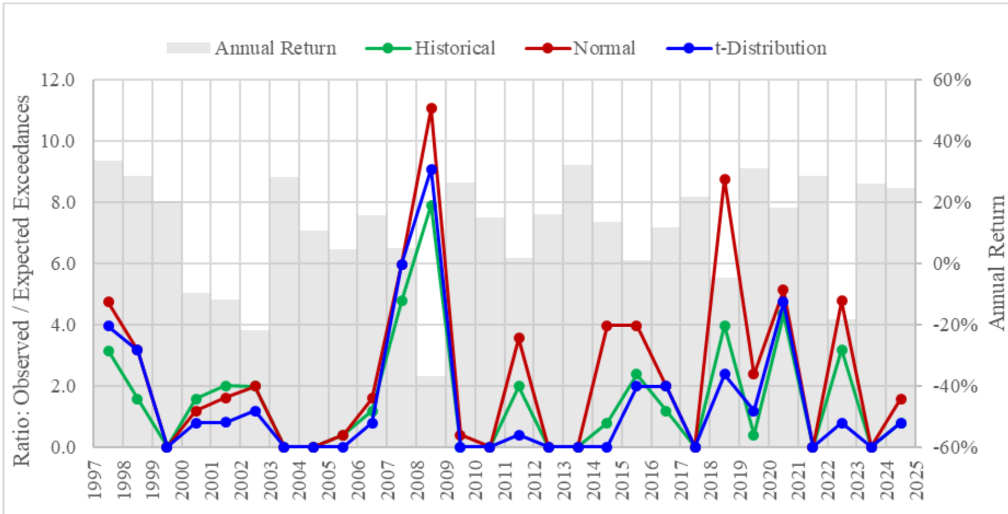


**Figure 9.** Annual rolling-window 99% VaR exceedance ratios and SPY annual returns. Annual observed-to-expected exceedance ratios for the 500-day rolling-window 99% VaR analysis. Gray bars represent annual SPY returns

**3.10** Practical economic implications of tail-risk underestimation

The tail-risk analysis results show that differences among statistical models are not only theoretical, but may also lead to materially different estimates of practical financial risk during stressed market conditions. In particular, the Expected Shortfall (ES) analysis shows that the normal distribution model consistently produces less conservative estimates of extreme downside losses compared to both empirical observations and heavy-tailed approaches. At the 99% confidence level, the normal-distribution ES estimate (-3.12%) is substantially smaller in magnitude than the empirical ES (-4.82%), while the t-distribution and EVT estimates remain considerably closer to the observed downside behavior. Additionally, the rolling-window backtesting analysis shows that the normal-distribution model produces observed 99% VaR exceedances at approximately 2.52 times the theoretically expected frequency, substantially higher than the rolling historical and t-distribution approaches. This underestimation of downside risk by the normal distribution, especially at deeper confidence levels, suggests that financial institutions relying on thin-tailed assumptions may systematically underestimate potential portfolio losses during periods of market stress. For large institutional portfolios, even relatively small percentage differences in estimated tail losses may correspond to substantial differences in liquidity requirements, leverage exposure, capital reserves, and realized drawdown severity during crisis periods.

The annual VaR and rolling-window backtesting analyses further show that model performance depends strongly on market regime and volatility conditions. During relatively stable market periods, the normal distribution often provides a reasonable approximation of moderate-tail behavior and may produce acceptable short-term calibration results. However, during stressed environments such as the 2008 financial crisis and the 2020 pandemic-related volatility period, exceedance ratios increased substantially across all models, particularly for the normal distribution. For example, during 2008, the annual rolling-window exceedance ratio of the normal model reached 11.07, compared to 7.91 for both the historical simulation and t-distribution models. Similarly, during 2018, the normal-distribution exceedance ratio reached 8.76, while the historical simulation and t-distribution models both remained at 3.98. These results suggest that the normal distribution increasingly underestimates downside risk during periods of elevated volatility and rapid market deterioration.

Overall, the results suggest that distributional model selection can materially influence practical estimates of portfolio risk, capital adequacy, leverage exposure, and potential drawdown severity, particularly during periods of elevated volatility and market stress. These findings have broader implications for practical financial risk management and regulatory stress testing, and they are consistent with modern market-risk frameworks that increasingly emphasize Expected Shortfall, stress testing, and backtesting performance because traditional normal-distribution assumptions may fail to capture heavy-tailed and time-varying market behavior during extreme events.

## 4. Conclusion

This study examined the distributional behavior and tail risk of SPY daily returns using multiple statistical models, including the normal distribution, t-distribution, GMM-2, and EVT. The results consistently show that SPY returns exhibit significant non-normal characteristics, including negative skewness, sharp central peaks, and heavy tails.

The normal distribution provides a reasonable approximation in moderate-tail regions and during stable market conditions, but increasingly underestimates downside risk at deeper confidence levels and during stressed market periods. The t-distribution provides the strongest overall balance between model simplicity and tail-risk representation among the tested parametric models, producing strong agreement with empirical VaR behavior and exceedance backtesting results. The GMM-2 model improves tail fitting relative to the normal distribution but tends to become overly conservative in deeper tails. EVT performs best for modeling rare and extreme losses, but is less suitable for moderate-risk regions.

The rolling-window sensitivity analysis further shows that the backtesting results remain relatively stable across different rolling-window lengths, suggesting that the main conclusions are not strongly dependent on the selected window lengths employed in this analysis. The annual rolling-window analysis also demonstrates that tail-risk model performance depends strongly on market regime and volatility conditions. In general, the normal distribution produces substantially larger exceedance ratios during stressed market periods, while the historical simulation and t-distribution models remain more consistent with the observed downside behavior.

Overall, the results demonstrate that tail-risk behavior is strongly time-dependent and clustered. Model performance varies substantially across different market environments, confidence levels, and volatility conditions. Consequently, no single model universally dominates across all market regimes and tail depths. This underscores the practical importance of adaptive model selection in financial risk management. Furthermore, the rolling-window analysis results reveal that accurate tail-risk forecasting depends not only on the selection of heavy-tailed distributions, but also on the ability to capture time-varying volatility and regime-dependent market behavior. These findings are broadly consistent with modern financial risk-management frameworks that increasingly emphasize stress testing, Expected Shortfall (ES), and dynamic backtesting under changing market conditions.

This study focuses only on SPY returns and therefore may not fully generalize to other asset classes or individual securities. In addition, SPY represents a dynamically changing market index whose constituent composition evolves over time, potentially introducing survivorship-related effects. Furthermore, VaR and ES results are sensitive to the selected sample period and the inclusion of major crisis events. Given that all tested models still significantly underestimated tail risk relative to theoretical exceedance expectations, future work should incorporate GARCH-type volatility models to better capture rapid volatility scaling, as well as multivariate portfolio analysis to evaluate these dynamic forecasting techniques across broader asset classes.  Future work should also explore regime-switching and model-selection frameworks that dynamically adjust between normal, heavy-tailed, and extreme-value approaches as market conditions evolve, potentially improving tail-risk calibration during periods of elevated volatility and market stress.

**Acknowledgements**

The author would like to acknowledge Mr. Y. Chen for providing guidance on the technical analysis and for reviewing the results and overall structure of the paper. The author would also like to sincerely thank the anonymous reviewer for the exceptionally thoughtful and constructive feedback throughout the review process. The reviewer’s comments substantially improved the technical depth, interpretation, and overall quality of this manuscript.

The author also acknowledges the use of ChatGPT for language editing, organizational suggestions, manuscript refinement, and improvements in structural flow and technical clarity, as well as Gemini for Python code debugging support. All AI-assisted content was reviewed, verified, and revised by the author, who retains full responsibility for the accuracy and integrity of the work.

## 5. References

1. Holden, S., Schrass, D., & Bogdan, M. (2024). Ownership of mutual funds and shareholder sentiment, 2024. *ICI Research Perspective, 30*(8). Investment Company Institute. https://doi.org/10.2139/ssrn.5279207

2. Dimson, E., Marsh, P., & Staunton, M. (2023). *Credit Suisse Global Investment Returns Yearbook 2023* (Summary ed.). Credit Suisse Research Institute.

https://www.credit-suisse.com/about-us/en/reports-research/global-investment-returns-yearbook.html

3. Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance,* 19(3), 425–442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

4. Jorion, P. (2007). *Value at risk: The new benchmark for managing financial risk* (3rd ed.). McGraw-Hill.

5. Fama, E. F. (1965). The behavior of stock-market prices. *The Journal of Business,* 38(1), 34–105. https://doi.org/10.1086/294743

6. Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance,* 1(2), 223–236. https://doi.org/10.1080/713665670

7. McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools* (Rev. ed.). Princeton University Press. https://doi.org/10.1515/9781400869021

8. Christoffersen, P. F. (2011). *Elements of financial risk management* (2nd ed.). Academic Press. https://doi.org/10.1016/C2010-0-66326-6

9. Hull, J. C. (2018). *Risk management and financial institutions* (5th ed.). Wiley. https://doi.org/10.1002/9781119448112

10. Albuquerque, R. (2012). Skewness in stock returns: Reconciling the evidence on firm versus aggregate returns. *The Review of Financial Studies,* 25(5), 1630–1673. https://doi.org/10.1093/rfs/hhr144